Best of both worlds?*

Simultaneous evaluation of researchers and their works

Ephrance Abu Ujum[1], Gangan Prathap[2], and Kuru Ratnavelu[1]

[1]Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia , ephrance@siswa.um.edu.my, kuru@um.edu.my
[2]CSIR National Institute for Interdisciplinary Science and Technology (NIIST), Council of Scientific and Industrial Research, Thiruvananthapuram, 695 019, Kerala, India, gp@niist.res.in

June 18, 2015

Abstract

This paper explores a dual score system that simultaneously evaluates the relative importance of researchers and their works. It is a modification of the CITEX algorithm recently described in Pal and Ruj (2015). Using available publication data for m author keywords (as a proxy for researchers) and n papers it is possible to construct a $m \times n$ author-paper feature matrix. This is further combined with citation data to construct a HITS-like algorithm that iteratively satisfies two criteria: first, *a good author is cited by good authors*, and second, *a good paper is cited by good authors*. Following Pal and Ruj, the resulting algorithm produces an author eigenscore and a paper eigenscore. The algorithm is tested on 213,530 citable publications listed under Thomson ISI's "*Information Science & Library Science*" JCR category from 1980–2012.

1 Introduction

Rankings provide an effective means to artificially assign order to the ever increasing volume of published research and researchers. The study and

*This work was supported by the University of Malaya High Impact Research Grant UM.C/625/1/HIR/MOHE/SC/13.

development of such work is increasingly trending towards what could be termed as *bibliometric analytics*, which we define here as[1] "key indicators derived from bibliometric data through mathematical or statistical analysis for the purpose of generating insight". In addition to information retrieval, bibliometric analytics focuses on discovering patterns specific to the data at hand in order to support decision-making or inference-related tasks. This paper is yet another step in this direction.

Specifically, this paper builds on recent work established by Pal and Ruj (2015) to simultaneously score research authors and papers by relative importance. The proposed algorithm, dubbed CITEX (CITation indEX), takes advantage of the many-to-many correspondence between a given set of authors and the papers they have collectively published. For this purpose, mappings between both sets can be formalized as linkages on a bipartite graph, hereon referred to as the author-paper (or author-document) network. In this sense, the cumulative advantage accrued by authors due to their papers and vice versa can be quantified using graph theoretic methods.

Furthermore, papers are interconnected through citation links; that is, a typical paper refers to previous works in order to acknowledge *relevance*[2] in addition to specifying its own *placement*[3] within the existing literature. The resulting paper citation network can thus be represented as a directed graph. Since the distribution of citation links varies from one paper to the next – usually in a highly skewed manner (Simon, 1955; de Solla Price, 1965; Price, 1976; Newman, 2009) – this can then be used as a basis to distinguish which papers are more prominently located than others. Several schemes have been proposed to exploit precisely this feature; i.e. scores are computed for each paper based on some discriminatory function of its connectivity (or how it is embedded within a structure of links) (Chen et al., 2007). These papers can then be ordered according to the computed scores to produce rankings. Such schemes are integral to information retrieval tasks on online databases, for example, Google Scholar, CiteSeerX, and Microsoft Academic Search.

CITEX extends this tradition by combining information from the author-paper network with the paper citation network to determine which authors

[1] We note that Bhatt and Martens (2009) and Rethlefsen and Aldrich (2013) used the term "*bibliometric analytics*" but have not provided a formal definition.

[2] In general, citation linkages are made to indicate *reaction* to past work rather than concrete *dependence*. Hence, the presence of citation linkages – that is, a link pointing from citing (referring) paper to cited (referred) paper – serves to describe intellectual flows in successive works, which in itself does not necessarily imply a flow of influence.

[3] This is a notion of the paper's *location* as opposed to its *position*. As with a citation count, the presence of a citation link does not explicitly convey whether it takes on the position of supporting or opposing the referred work.

and which papers stands out more than others. The development of such algorithms are important to explore alternative means of assembling bibliometric indicators (and their derived rankings) through purposeful integration of available information. CITEX is interesting in its construction because it provides a coupled dual score system: a relative importance score for authors and another for papers, hence, the relative standings of knowledge creators and the results of their labours can be determined within a single framework. Simply put, CITEX asserts that: (1) good authors are either highly prolific with, or are highly cited by good authors; and, (2) good papers share the same authors with, or are cited by good papers.

This paper is organized as follows. We provide an in-depth discussion on the construction of the CITEX algorithm in Section 2 and a critique is offered in Section 3. Our proposed modification, hereon referred to as the CAPS (Coupled Author-Paper Scoring) algorithm, is then described in detail in Section 4. To provide a point of comparison, both algorithms are tested on a real world dataset in Section 5. This consists of 200,000+ ISI-cited papers published from 1980-2012 listed under the Journal Citation Reports subject category of "*Information Science & Library Science*". The paper is concluded in Section 6.

2 The CITEX algorithm

Suppose we are presented with a corpus consisting of m authors and n papers. Furthermore, suppose that from this corpus, we are able to extract the binary $m \times n$ author-paper feature matrix, M, and binary $n \times n$ citation matrix, C. Let an entry $M_{ij} = 1$ denote that author i on the i-th row of M has (co)authored paper j on the j-th column of M ($M_{ij} = 0$ otherwise). This implies that row sums of M correspond to total papers published by each author. Column sums of M correspond to total authors for each paper. A column-normalized version of M (with the same dimensions) can be constructed so that *authorship share* of author i to paper j is divided equally as $W_{ij} = M_{ij}/\sum_i M_{ij}$.

In a similar way, let $C_{ij} = 1$ denote that cited paper j on the j-th column of C receives a citation from a citing paper i on the i-th row of C ($C_{ij} = 0$ otherwise). Additionally, we require that C contains no self-citations ($C_{ii} = 0$). Given an extreme case where $C = 0_{n \times n}$, Pal and Ruj define the CITEX paper and author scores as $y_j = \sum_{i=1}^{m} M_{ij} x_i$ and $x_i = \sum_{j=1}^{n} W_{ij} y_j$, respectively. These expressions are written in matrix form as $y \leftarrow M^T x$ and $x \leftarrow Wy$. This captures the notion that the y-score for paper j depends on the relative importance of its authors, while the x-score for author i depends on her authorship share (W_{ij}) for each paper j multiplied by its corresponding score y_j.

A complete description however requires the inclusion of citation features. Since this must reduce to the case of a zero citation matrix, Pal and Ruj achieve this by the inclusion of a $(I + C^T)$ term (which is equivalent to adding in paper self-citations to C). Since $y \leftarrow M^T W y$ and $x \leftarrow W M^T x$, then for the k-th recursion:

$$x^{(k)} = W(I + C^T)M^T x^{(k-1)} \tag{1}$$

$$y^{(k)} = (I + C^T)M^T W y^{(k-1)} \tag{2}$$

is one such possible choice. By induction, we obtain:

$$x^{(k)} = [W(I + C^T)M^T]^k x^{(0)} \tag{3}$$

$$y^{(k)} = [(I + C^T)M^T W]^k y^{(0)} \tag{4}$$

For initial guess vectors, Pal and Ruj use $x^{(0)} = 1_{m\times 1}$ and $y^{(0)} = 1_{n\times 1}$. Supposing $P = W(I + C^T)M^T$, so that $x^{(k)} = P^k x^{(0)}$, then:

$$x^{(k+1)} = PP^k x^{(0)} = Px^{(k)} \tag{5}$$

If the distance between two x score vectors is $\|x^{(k+1)} - x^{(k)}\| < \epsilon$ then convergence is met relative to tolerance ϵ (Franceschet, 2011). Since P is a nonnegative matrix with dimensions $n \times n$ and $x^{(0)} > 0$, then in accordance with the Perron-Frobenius theorem[4], the x scores become stationary as $k \to \infty$, thus satisfying $Px^* = x^*$ (Perron, 1907; Frobenius, 1912). A similar argument is applicable for y by setting $Q = (I + C^T)M^T W$.

There are other algorithms that combine author and paper features. One notable example is the Co-Ranking framework proposed in Zhou et al. (2007). This approach uses a PageRank-based model on a bipartite co-authorship/paper citation network, whereby two intra-class random walks allow traversal strictly between one class of nodes, while an inter-class random walk allows jumps between networks. The stationary probabilities for author nodes and paper nodes are computed by coupling the random walks (assuming the status of researchers and the work they produce are mutually reinforced). The resulting algorithm yields improvements compared to when applying PageRank on either feature (network) in isolation, although at the expense of introducing three additional adjustable parameters to the usual

[4] In particular, given that $Px = cx$ and $c = 1$ is the largest eigenvalue, then $P^k x^{(0)}$ converge to a vector x^* (in the same direction as x) as $k \to \infty$.

one-parameter PageRank[5]. CITEX adds an interesting twist to the current literature since, unlike PageRank, it does not depend on any adjustable parameters.

3 Expected behaviour and blindspots

Since the performance of a data mining algorithm depends on its design (Jahne, 2000; Balakin, 2010), it is useful to determine precisely what features are emphasized by CITEX in order to anticipate the qualitative aspects of the ranking it will necessarily produce. In particular, we are interested in the conditions that maximize a given score since the highest percentile is designed to correspond to the topmost ranks. Specific to the CITEX author score, Equation 1 can be expanded as:

$$x^{(k)} = WM^Tx^{(k-1)} + WC^TM^Tx^{(k-1)} \tag{6}$$

$$x_i^{(k)} = \sum_{a=1}^{m}\sum_{p=1}^{n} W_{ip}M_{ap}x_a^{(k-1)} + \sum_{a=1}^{m}\sum_{p_1,p_2=1}^{n} W_{ip_1}C_{p_2p_1}M_{ap_2}x_a^{(k-1)} \tag{7}$$

The first term on the right hand side of Equation 7 captures the *cumulative authorship share* of author i with author a. This term is positively biased towards author i if she is prolific (adjusting for authorship share), and more so if she collaborates frequently with "good authors" (those with high x-scores). This includes the case where $a = i$, so that if the cumulative authorship share of i herself is significantly large, then $x_i^{(k)} \sim x_i^{(k-1)} \sum_{p=1}^{n} W_{ip}$.

As for the second term, a citation from paper $p_2 \to p_1$ corresponds to an author citation from $a \to i$ fractionalized by W_{ip_1}. Hence, this term increases the larger the number of citations from $a \to i$, the larger the authorship share for each paper authored by i (for which credit is minimally split), and the larger the x-score of i's citing authors. Put together, *CITEX defines a good author as one who publishes frequently with good authors, and is even more so if he/she is cited by good authors.*

A similar analysis can be done for the CITEX paper score as given in Equation 2:

$$y^{(k)} = M^TWy^{(k-1)} + C^TM^TWy^{(k-1)} \tag{8}$$

$$y_j^{(k)} = \sum_{a=1}^{m}\sum_{p=1}^{n} M_{aj}W_{ap}y_p^{(k-1)} + \sum_{a_1,a_2=1}^{m}\sum_{p=1}^{n} C_{pj}M_{a_1j}W_{a_2p}y_p^{(k-1)} \tag{9}$$

[5] We are referring to the *damping parameter* originally described in Brin and Page (1998). The interested reader is referred to Langville and Meyer (2006) and Chen et al. (2007) for an in-depth discussion on the PageRank algorithm.

From the right hand side of Equation 9, we see again that CITEX defines relative importance in terms of two components; the first term captures publication features while the second term captures citation features.

For the first term, we see that paper j receives fractional y-score contributions for each author a appearing in both papers j and p. Essentially, $M_{aj}W_{ap}$ is an *author similarity* term, hence, this part of the equation increases for papers that share the same authors. This term will also increase the larger the y-score for each "similar author" paper p (relative to j) and whenever $W_{ap} \to 1$. For the case of an author i with a significantly large number of papers, we could end up with $y_j^{(k)} \sim y_j^{(k-1)} \sum_{p=1}^{n} W_{ip}$.

For the second term, we see that y_j depends on the sum of y-scores from each paper that cites it, p ("good papers" have high y-score). With some rearranging, the second term also contains the product $W_{a_2p}C_{pj}M_{a_1j}$. This means that the y-score of paper j depends on the sum of fractionalized citations from all citing papers p (i.e. $\sum_{p=1}^{n} W_{a_2p}C_{pj}$). Combining this with the effect from the first term of Equation 9, we surmise that *CITEX defines a good paper as one with high author similarity with good papers, and is even more so if it is cited by good papers.*

Based on our analysis, we have determined two quirks with the original formulation of CITEX. These are:

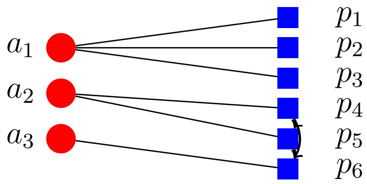
1. $x_i^{(k)} \sim x_i^{(k-1)} \sum_{p=1}^{n} W_{ip}$: the CITEX author score for an author i can increase from being highly prolific, and more so if he/she tends to coauthor in small teams. This allows for the case of an extremely prolific solo author to be over-represented by the algorithm. He or she may not even need a boost from citation count (from good authors or otherwise) in order to obtain a high CITEX author score.

2. $y_j^{(k)} \sim y_j^{(k-1)} \sum_{p=1}^{n} W_{ip}$: the CITEX paper score can increase just by having the same author list repeat over a significant fraction of the collection, with this effect becoming more pronounced if the listing tends to be short. Similarly, such cases can be over-represented by CITEX without a boost from citation count (from good papers or otherwise).

To illustrate the potential problems associated with these quirks, we construct two toy calculations analogous to those posed in Pal and Ruj (2015). These are as shown in Figure 1 and Figure 2.

As a result of the quirks highlighted in Figure 1 and Figure 2, we can expect that author and paper rankings generated by CITEX will suffer from specificity issues since extreme publication and citation traits are mixed together. The task of this paper is to propose a more elegant variation of the CITEX algorithm that addresses the above mentioned issues.

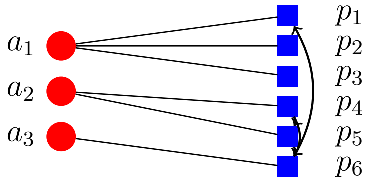
$$M = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 1: Problem 1 — Hypothetical case of one prolific solo author with no citations. CITEX gives $x = [0.333, 0.333, 0.333]$ and $y = [0.143, 0.143, 0.143, 0.095, 0.191, 0.285]$. Hence, all three authors are ranked equally even though there are stark qualitative differences between their publication and citation patterns. Understandably, paper p_6 has the highest score followed by p_5 due to the number of citations they receive compared to no citations for the other papers. Oddly, p_4 is ranked lower than papers p_1, p_2 and p_3 despite being authored by author a_2 who has one citation more than a_1 (via p_5).

$$M = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 2: Problem 2 — The effect of high author similarity with good papers. The setup in this diagram is similar to Figure 1 with one additional citation link added from paper p_6 to p_1. CITEX gives $x = [0.521, 0.214, 0.214]$ and $y = [0.243, 0.175, 0.175, 0.068, 0.136, 0.203]$. Author a_1 leads by author score followed by a tie between a_2 and a_3, despite the absence of (co)author self-citations to a_1 (note that a_2 has one author self-citation via $p_4 \rightarrow p_5$). Paper p_1 is ranked highest despite having only one citation because it is cited by a good paper (p_6). Due to the way paper scores are propagated in CITEX, papers p_2 and p_3 also receive high scores just by having high author similarity with paper p_1.

4 An improved Coupled Author-Paper Scoring algorithm

As highlighted in Section 3, CITEX has a built-in tendency to produce a rank ordering that gives undesired priority to highly productive authors (even if they are relatively uninfluential), in addition to assigning high relative importance to papers associated to highly prolific authors (overriding the citation impact of other papers).

To circumvent these issues, we propose dropping the self-citation term $(I + C^T)$ in Equations 1 and 2, and replace the M matrices with W matrices to ensure conservation of citation count when switching from the paper citation network to the author citation network (inter-author citations are fractionalized). This results in the following set of equations which defines our **Coupled Author-Paper Scoring** (**CAPS**) algorithm:

$$x^{(k)} = WC^T W^T x^{(k-1)} \tag{10}$$

$$y^{(k)} = C^T W^T x^{(k)} \tag{11}$$

Following previous conventions (Kleinberg, 1999; Pal and Ruj, 2015), we start with an initial guess vector (specifically, $x^{(0)} = 1_{m\times 1}$ and $y^{(0)} = 1_{n\times 1}$) and determine the values of scores iteratively (i.e. iterate $k \geq 1$ until convergence is achieved for a given tolerance level).

Equation 10 quantifies the criterion that "*a good author is cited by good authors*". Equation 11 quantifies the criterion that "*a good paper is cited by good authors*". The equations above provide a self-consistent basis for repeated improvement (Easley and Kleinberg, 2010, pp. 355–356). This can be seen by writing $L = WC$:

$$x^{(k)} = WL^T x^{(k-1)} = Wy^{(k-1)} \tag{12}$$

$$y^{(k)} = L^T x^{(k)} \tag{13}$$

Hence, *a good author has good papers that are cited by good authors who have good papers* and so on. The $m \times n$ matrix L has entries $(L)_{ij} = \sum_{p=1}^{n} W_{ip} C_{pj}$ which correspond to the cumulative fractional citations made by citing author i through papers p (if authored by i) to some cited paper j. Essentially, L encodes the *author-paper citation matrix.*

Entries of the $m \times m$ matrix product WL^T in Equation 12 corresponds to the cumulative fractional citations received by authors in row i from authors in column a. This is because $(WL^T)_{ia} = \sum_{p_1,p_2=1}^{n} W_{ap_2} C_{p_2 p_1} W_{ip_1}$ signifies that author a in paper p_2 cites paper p_1 which is (co)authored by i. The sum over all possible papers p_1 serves to aggregate all fractional citations received by author i from author j. WL^T is thus the (fractional) *author citation matrix.*

In effect, we find that the author score defined in Equation 12 therefore corresponds to $x_i^{(k)} = \sum_{a=1}^{m} \sum_{p_1,p_2=1}^{n} W_{ap_2} C_{p_2 p_1} W_{ip_1} x_a^{(k-1)}$. Therefore, the author score for author i is proportional to the cumulative author citations received as well as the score of the citing authors. This captures the intuition that *authors promote each other through their published works.* Similarly, Equation 13 implies that the paper score for paper j is $y_j^{(k)} = \sum_{i=1}^{m} L_{ij} x_i^{(k)}$. This quantifies the relationship that *the relative importance of a paper depends on the authority its citing authors.*

5 Empirical test

We test the CITEX and CAPS algorithm on papers published under the Thomson ISI *Journal Citation Reports* (JCR) subject category of "*Information Science & Library Science*" (LIS) from the years 1980 up to 2012 inclusive. This dataset consists of 213,530 papers, 471,191 total inter-paper citations, and 73,597 author keywords. We do not conduct author or bibliographic reference disambiguation in order to assess the output quality of CAPS and CITEX when used with minimal data preprocessing.

5.1 Authors

The output of a ranking scheme depends on how it scores selected features that are present (or absent) for each datum relative to the rest of the dataset. In general, it is difficult to determine the performance of the underlying scoring algorithm when there is no ground truth to base such judgements. In cases like this, the most sensible thing to do is to speak of the properties of the scores generated by the algorithm of interest, and whether the rankings generated show reasonable agreement with known methods and observations.

In this respect, the distribution of author scores for CAPS and CITEX exhibit a reasonably high Spearman rank correlation coefficient (ρ) with h-index score ($p < 0.01$): specifically, 0.77 and 0.69 for CAPS and CITEX, respectively. The h-index (Hirsch, 2005) provides a useful comparison to CAPS and CITEX as it too combines publication and citation traits together. However, unlike CAPS (and to a lesser extent, CITEX), the h-index is not designed to differentiate whether a citation is received from a relatively "good" paper (author) or otherwise, hence some disparity in the resulting ranking is to be expected. This can be seen in Table 1.

Since CAPS and CITEX are also positively correlated with $\rho = 0.85$

Table 1: Top 25 (out of 73,597) authors by publication count, citation count, CAPS author score, CITEX author score, and h-index, respectively. Note that h-index values in columns denoted by h are computed using available data (only ISI papers indexed under LIS JCR category from 1980-2012). Note the usage of ordinal ranking for the h-index column.

		Pubs.		Times Cited		CAPS		CITEX		h-index
Rank	h	Author key	h	Author key	h	Author key	h	Author key	h	Author key
1	5	rogers.m	5	davis.fd	21	egghe.l	5	rogers.m	30	glanzel.w
2	0	cassada.j	29	benbasat.i	26	leydesdorff.l	0	cassada.j	29	bates.dw
3	0	klett.re	12	venkatesh.v	23	rousseau.r	0	klett.re	29	benbasat.i
4	1	ramsdell.k	29	bates.dw	30	glanzel.w	1	ramsdell.k	28	garfield.e
5	1	christian.g	1	pawlak.z	24	thelwall.m	1	christian.g	26	leydesdorff.l
6	0	vicarel.ja	19	straub.dw	16	burrell.ql	0	vicarel.ja	25	schubert.a
7	2	hoffert.b	30	glanzel.w	25	schubert.a	2	hoffert.b	25	spink.a
8	1	sutton.j	1	gruber.tr	17	ingwersen.p	1	sutton.j	24	grover.v
9	1	sutton.jc	25	spink.a	17	bar-ilan.j	24	thelwall.m	24	moed.hf
10	1	bigelow.d	14	salton.g	22	braun.t	21	egghe.l	24	thelwall.m
11	1	stevens.n	3	furnas.gw	18	cronin.b	30	glanzel.w	23	rousseau.r
12	0	zlendich.j	24	grover.v	17	van.raan.afj	26	leydesdorff.l	22	braun.t
13	0	fairchild.ca	3	deerwester.s	16	white.hd	1	sutton.jc	21	egghe.l
14	1	pearl.n	3	dumais.st	12	jacso.p	2	decandido.ga	20	willett.p
15	0	richard.o	2	landauer.tk	24	moed.hf	23	rousseau.r	19	ford.n
16	2	gordon.rs	8	buckley.c	15	vinkler.p	3	stlifer.e	19	saracevic.t
17	0	maccann.d	25	schubert.a	16	small.h	1	bigelow.d	19	smaglik.p
18	0	lombardo.d	5	morris.mg	18	mccain.kw	25	schubert.a	19	straub.dw
19	1	williamson.ga	1	harshman.r	16	bornmann.l	2	rawlinson.n	18	bates.mj
20	5	butler.t	7	todd.pa	28	garfield.e	0	davidson.a	18	chen.hc
21	1	raiteri.s	9	karahanna.e	9	pao.ml	0	de.baron.fhk	18	cronin.b
22	1	gillespie.t	26	leydesdorff.l	15	vaughan.l	0	elizabeth.p	18	dennis.ar
23	1	campbell.p	18	zmud.rw	6	rao.ikr	1	furlong.cw	18	lyytinen.k
24	3	burns.a	14	gefen.d	15	daniel.hd	0	hammett.d	18	mccain.kw
25	2	wyatt.n	19	saracevic.t	15	oppenheim.c	0	hemingway.h	18	zmud.rw

($p < 0.01$), we can expect that the h-index distribution for top N ranks by CAPS and CITEX score to resemble each other for increasingly large N. For the top $N = 25$ ranks, $\mu_{\text{CAPS}}(h) = 18.44$ while $\mu_{\text{CITEX}}(h) = 6.76$. For $N = 250$ the mean h-index values are 8.03 and 7.03, while for $N = 2500$ we obtain 3.32 and 3.36 for CAPS and CITEX, respectively. Ideally, the top percentile of any ranking should correspond to an easily interpreted ordering by quality, hence in this sense, CAPS improves on the CITEX author ranking (since the top ranks tend to correspond to high h-index values).

Incidentally, the top ranked author by CITEX (Rogers, with a score of 3.37×10^{-5}) corresponds to 83.6% of the entire CITEX author score distribution. Together with Cassada (author score $= 2.39 \times 10^{-14}$), both authors take up a shocking 96% of total scores. Over the entire list of authors, this corresponds to a Gini coefficient[6] of 0.9999. In contrast, 20% (14,719) of top scoring authors according to the CAPS algorithm accounts for approximately 99.96% of the scores (corresponds to a Gini coefficient of 0.9891). This implies that the difference between CAPS author scores for adjacent ranks becomes progressively smaller as we go down the ranks. This is exaggerated to a greater extreme in CITEX.

Interestingly, the Gini coefficients for fractional publication count and fractional citation count of authors in the LIS dataset are 0.7744 and 0.8715, respectively. Furthermore, 20% of top authors account for 81.4% of the total fractional publications as well as 90% of the total fractional citations. While these values are characteristic of high levels of inequailty, they are quite tame compared to the level of inequality implied by CAPS. The presence of such extreme levels of inequality suggests a vast differential in the ability of LIS researchers to capitalize the resources, technical skills, and opportunities at their disposal (Shockley, 1957).

5.2 Papers

The top 25 ranking by citation count, CAPS paper score, and CITEX paper score is as displayed in Table 2. The topmost ranks of CITEX are populated by papers sharing the same high-scoring author (Rogers). Looking beyond the top 25 ranks, we find that with the exception of papers at ranks 7 to 12, the first 3819 positions are papers authored by Rogers, while the next 2610 positions (ranks 3820−6429) are papers authored by Cassada. Hence, CITEX tends to over-represent the importance of papers from the same highly scored author even if these do not correspond to "high impact" works or works that impact "high impact works". This is precisely the effect we described in Section 3.

As we have seen in the case of authors, the paper citation data shows high inequality since the top 10% of cited papers accounts for nearly 88.8% of total citations. This is expected since only a fraction of papers are cited and each of these papers receives additional citation in-links at a rate proportional to their current number of citation in-links. This suggests that the citation distribution is governed by a cumulative advantage/preferential attachment process whereby the *rich get richer* (Price, 1976; Barabási et al., 1999).

[6] The Gini coefficient is a measure of statistical dispersion typically used to measure the level of inequality in a given sample. For a sample of size n ordered such that $x_i \leq x_{i+1}$, it is given by $G = \frac{2\sum_{i=1}^{n} ix_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n}$. A Gini coefficient of 1 indicates maximal inequality whereby the total score is associated to only one element in the sample while the remainder of the sample contributes nothing to the total score. A Gini coefficient of 0 indicates perfect equality whereby the total score is distributed equally among all elements in the sample.

Table 2: Top 25 (out of 213,530) papers by citation count, CAPS paper score, and CITEX paper score. Papers are identified by publication year, followed by source journal abbreviation, volume, page number, and first author. Source journal abbreviations are listed in Table 3. TC designates the times cited for papers as reported by ISI in 2012. The Spearman rank correlation coefficients ($p < 0.01$) over all papers are: $\rho(C_1, C_2) = 0.87$, $\rho(C_1, C_3) = 0.17$, and $\rho(C_2, C_3) = 0.26$. CAPS appears in better agreement with citation count than CITEX.

	Citation count (C_1)		CAPS (C_2)		CITEX (C_3)	
Rank	Paper	TC	Paper	TC	Paper	TC
1	1982/IJCIS/11/341/pawlak	3319	2006/SCI/69/121/egghe	105	1995/LJ/120/113/rogers	1
2	1989/MISQ/13/319/davis	3251	1990/JIS/16/17/egghe	61	1995/LJ/120/119/rogers	1
3	1993/KA/5/199/gruber	2618	2006/SCI/69/131/egghe	250	1995/LJ/120/130/rogers	1
4	1990/JASIS/41/391/deerwester	2150	1998/JD/54/236/ingwersen	199	1995/LJ/120/187/rogers	1
5	1980/PAL/14/130/porter	1653	2005/S/19/8/braun	86	1995/LJ/120/213/rogers	1
6	2003/MISQ/27/425/venkatesh	1534	2003/JASIST/54/550/ahlgren	123	1996/LJ/121/100/rogers	1
7	1988/IPM/24/513/salton	1449	2006/SCI/69/169/braun	127	2011/LJ/136/30/fox	0
8	2001/MISQ/25/107/alavi	1075	2006/SCI/67/491/van.raan	177	2007/LJ/132/36/albanese	4
9	1995/ISR/6/144/taylor	1021	1999/JD/55/577/smith	93	1993/LJ/118/32/berry	2
10	2003/JMIS/19/9/delone	772	2001/JASIST/52/1157/thelwall	94	1989/LJ/114/18/decandido	1
11	2004/MISQ/28/75/hevner	724	1985/JD/41/173/egghe	48	1989/LJ/114/57/decandido	0
12	1995/MISQ/19/189/compeau	684	1992/IPM/28/201/egghe	41	1995/LJ/120/12/stlifer	1
13	2003/MISQ/27/51/gefen	677	1989/SCI/16/3/schubert	165	1992/LJ/117/52/rogers	0
14	2000/ISR/11/342/venkatesh	596	1997/JD/53/404/almind	163	2000/LJ/125/91/rogers	0
15	1999/MISQ/23/67/klein	569	2001/SCI/50/65/bjorneborn	93	2005/LJ/130/172/rogers	0
16	2000/MISQ/24/169/bharadwaj	568	1986/SCI/9/281/schubert	162	2006/LJ/131/114/rogers	0
17	1992/MISQ/16/227/adams	542	2006/SCI/67/315/glanzel	88	2006/LJ/131/114/rogers	0
18	1995/MISQ/19/213/goodhue	540	2006/SCI/69/161/banks	60	2006/LJ/131/123/rogers	0
19	1987/MISQ/11/369/benbasat	526	1996/SCI/36/97/egghe	31	2006/LJ/131/123/rogers	0
20	1999/MISQ/23/183/karahanna	513	1991/JASIS/42/479/egghe	29	2007/LJ/132/132/rogers	1
21	1999/JAMIA/6/313/bates	497	2003/SCI/56/357/glanzel	82	2007/LJ/132/132/rogers	1
22	1988/MISQ/12/259/doll	477	1986/SCI/9/103/leydesdorff	46	2007/LJ/132/171/rogers	0
23	1999/IJGIS/13/143/stockwell	475	2001/SCI/50/7/bar-ilan	64	2007/LJ/132/96/rogers	0
24	2000/MISQ/24/115/venkatesh	475	2002/JASIST/53/995/thelwall	72	2006/LJ/131/27/rogers	1
25	2003/ISR/14/189/chin	472	1996/JIS/22/165/egghe	24	2003/LJ/128/40/rogers	2

In contrast, the CAPS paper score possesses a Gini coefficient of 0.9912, while CITEX has a slightly lower value of 0.9785. This implies that both methods exhibit large score differentials only between the topmost ranks. For CAPS paper score, this can be traced to the fact that 81.2% of the lowest scoring population has a score of exactly zero (76% of papers in the study data have zero citations[7]). The reason for this is that the coupling

[7] The LIS dataset consists of 103,768 papers from *Library Journal* ($\sim$ 48.6% of total). This is nearly 14 times larger than the the 2^{nd} largest contributor, i.e. *Scientist.* While this seems excessively high, consider that only 1.9% of papers from *Library Journal* contributes 1% of non-zero citations in the LIS dataset (from a total of 471,191 citations for 213,530 papers). In comparison, *Scientometrics* is only the 6^{th} largest contributor to the dataset

Table 3: Journal composition in top 100 ranks by algorithm.

Citation count			CAPS			CITEX		
MISQ	mis.quart	43	SCI	scientometrics	33	LJ	libr.j	100
ISR	inform.syst.res	14	JASIS	j.am.soc.inform.sci	17			
JAMIA	j.am.med.inform.assn	9	JASIST	j.am.soc.inf.sci.tec	14			
JASIS	j.am.soc.inform.sci	6	JD	j.doc	12			
JD	j.doc	3	JIS	j.inform.sci	9			
IPM	inform.process.manag	3	IPM	inform.process.manag	6			
JMIS	j.manage.inform.syst	3	JI	j.informetr	5			
IJCIS	int.j.comput.inf.sci	2	ARIS	annu.rev.inform.sci	2			
IM	inform.manage	2	SSI	soc.sci.inform	1			
IJGIS	int.j.geogr.inf.sci	2	S	scientist	1			
SCI	scientometrics	2						
ARIS	annu.rev.inform.sci	1						
CJIS	can.j.inform.sci	1						
EJIS	eur.j.inform.syst	1						
GIQ	gov.inform.q	1						
IJGIS	int.j.geogr.inf.syst	1						
IMA	inform.manage-amster	1						
JASIST	j.am.soc.inf.sci.tec	1						
JIS	j.inf.sci	1						
KA	knowl.acquis	1						
OR	online.rev	1						
PAL	program-autom.libr	1						

of both author features and paper features places strict limits on the size of the non-zero scoring population. On the other hand, the CITEX paper score has no zero scoring population (due to the presence of artificial paper self-citations). The extremely high Gini coefficients for both CITEX and CAPS[8] implies that we can only reasonably differentiate a small fraction of the dataset corresponding to top scoring papers that coincide with top scoring authors.

A quick glance at top scoring papers listed in the "Citation count" column of Table 2 reveals that these mostly correspond to informatics papers rather than informetrics. Contrast this with the listing shown in the "CAPS" column where the emphasis is more towards informetrics papers instead. The reason for this is that the CAPS algorithm takes into account authorship features when scoring papers, which are not accounted for in a simple citation count. Since informetrics authors are highlighted in Table 1, it follows that informetrics papers are also highlighted in Table 2. Table 3 provides a listing of journals in the top 100 ranks. This provides some indication of the

with 3100 papers (1.5% of total LIS papers) yet contributes a total of 29792 citations (6.3% from LIS total) making it the 4^{th} largest contributor citation-wise. For reference, the largest citation counts are attributed to *MIS Quarterly*, *J AM MED INFORM ASSN*, and *J AM SOC INFORM SCI* with 55736, 30470, 30317 citations, respectively.

[8] For CITEX, 1% of the top scoring population accounts for 50.1% of the total score, while 2% accounts for 91%. In comparison, CAPS has 1% and 2% of the top scoring population accounting for 88% and 97% of total scores, respectively.

research field predominantly featured by each method.

6 Conclusion

In this paper we have constructed a modified version of the CITEX algorithm originally introduced by Pal and Ruj (2015). This algorithm was designed to assign relative importance scores to papers and authors by taking into account data from both entities simultaneously. Conventional methods like citation count and PageRank, for example, cannot do so without appropriate modification. The modification of CITEX which we propose, dubbed the CAPS (Coupled Author-Paper Scoring) algorithm, is designed to address some of the weaknesses of Pal and Ruj's original algorithm which we described in Section 3 (essentially, the shortcomings can be traced to artificially introduced self-citations on the paper-level).

Using a real dataset (ISI papers published from $1980 - 2012$ in the JCR subject category of "*Information Science & Library Science*"), we show that our proposed modifications outperforms CITEX in identifying important authors and papers. However, the CAPS algorithm appears to suffer from high inequality in the resulting score distributions as indicated by an extremely high Gini coefficient (~ 0.99). The inequality is similarly pronounced in CITEX. This implies that both CAPS and CITEX generate extreme prejudice in the allocation of scores to the top scoring minority.

However, this is not necessarily a bad thing. By design, CAPS allocates high scores to authors and papers associated to instances where the likelihood of future success (an increase in publication count or citation count) is proportional to previous success. Hence, CAPS can be used to highlight parts of the data attributed to the "*rich get richer*" effect. In contrast, CITEX rewards high scores for authors lying at the tail of the publication productivity distribution, and by association, rewards high scores for papers published by such authors irrespective of the relative importance of their papers within the paper citation network. In this sense, CITEX is useful to find instances where high productivity is mismatched with low impact.

While bibliometric analytic algorithms such as CITEX or CAPS, or even bibliometric adaptations of website ranking algorithms such as HITS or PageRank can prove useful in identifying what is important in a given dataset, it is crucial to be aware of the limitations and subtleties of such methods. Each method finds exactly what it is designed to seek and since it is hard to account for, let alone anticipate every relevant feature or contingency, we must concede that the rankings produced are themselves only facets of the underlying organization in the data. Hence, bibliometric analytic algorithms should be used first and foremost to guide decisions on where to look deeper (i.e. to construct recommendation engines), and if necessary, used with extreme caution when drawing inferences on the relative standing of bibliometric entities.

References

Balakin, K. V. (2010). *Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery.* John Wiley & Sons, Hoboken, New Jersey, USA.

Barabási, A., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1):173–187.

Bhatt, A. and Martens, B. (2009). THE TOPICS OF CAAD: AN EVOLUTIONARY PERSPECTIVE. In Tidafi, T. and Dorta, T., editors, *Joining Languages, Cultures and Visions: CAAD Futures 2009*, Proceedings of the 13th International CAAD Futures Conference, Montréal. Les Presses de l'Université de Montréal.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117.

Chen, P., Xie, H., Maslov, S., and Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1):8–15.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets.* Cambridge Univ Press.

Franceschet, M. (2011). Pagerank: Standing on the shoulders of giants. *Communications of the ACM*, 54(6):92–101.

Frobenius, G. F. (1912). Über Matrizen aus nicht negativen Elementen. *Königliche Akademie der Wissenschaften*, pages 456–477.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569.

Jahne, B. (2000). *Computer Vision and Applications: A Guide for Students and Practitioners*. Academic Press, San Diego, CA, USA.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Langville, A. N. and Meyer, C. D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, New Jersey, USA.

Newman, M. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, 86(6):68001.

Pal, A. and Ruj, S. (2015). CITEX: A new citation index to measure the relative importance of authors and papers in scientific publications. *arXiv preprint arXiv:1501.04894*.

Perron, O. (1907). Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263.

Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.

Rethlefsen, M. L. and Aldrich, A. M. (2013). Environmental health citation patterns: mapping the literature 2008–2010. *Journal of the Medical Library Association: JMLA*, 101(1):47.

Shockley, W. (1957). On the Statistics of Individual Variations of Productivity in Research Laboratories. *Proceedings of the IRE*, 45(3):279–290.

Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, pages 425–440.

Zhou, D., Orshanskiy, S. A., Zha, H., and Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. In *Seventh IEEE International Conference on Data Mining, 2007 (ICDM 2007)*, pages 739–744. IEEE.